Descripción de recursos multimedia georreferenciados

A. Beltran Fonollosa[(1)], *C. Granell Canut*[(1)] *y J. Huerta Guijarro*[(1)]

[(1)] Institute of New Imaging Technologies (INIT). Universidad Jaume I de Castellón, Avda. de Vicente Sos Baynat s/n, E-12071 Castellón, {arturo.beltran, carlos.granell, huerta}@uji.es.

RESUMEN

La información geográfica juega un papel fundamental en la sociedad actual, y el interés de los usuarios por ella crece día a día. Sin embargo, aún resulta demasiado complicado encontrar contenidos geográficos que sean relevantes (actualizados, de calidad y veraces), pese a los esfuerzos realizados en generar grandes catálogos de metadatos.

Para hacer que la información esté disponible a nivel global y llegue fácilmente al mayor número de personas posible resulta esencial organizar, publicitar y facilitar el acceso a dicha información. Y para que esto sea posible, es decir, para que un recurso sea encontrado como resultado de una búsqueda es necesario describirlo según sus propiedades. Es en este contexto donde los metadatos cobran sentido y se convierten en la pieza central de cualquier sistema de información.

Con el objetivo de conseguir descripciones de los recursos, se analizaron diferentes herramientas de extracción de metadatos. Se consideró como la propuesta más interesante la del proyecto Apache Tika. Resulta una interesante plataforma común de extracción de metadatos para recursos multimedia, el problema es que no soporta formatos de información geoespacial. En consecuencia, se buscaron y evaluaron diferentes plataformas comunes de acceso a datos geográficos. Entre las diferentes opciones analizadas, destaca la plataforma FDO (FDO Data Access Technology) desarrollada por OSGeo.

En este trabajo se pretende conseguir una plataforma que permita obtener información y acceder de forma lo más homogénea posible a recursos heterogéneos para poder extraer metadatos, con especial interés en los recursos geoespaciales. Este sería el primer paso para la descripción de los recursos, mediante la cual, posteriormente se podrá abordar la publicación, ya sea mediante la indexación respecto a sus características o la inclusión en un servidor de catálogo.

Palabras clave: *Descripción, Metadatos, Multimedia, Información geoespacial*

INTRODUCCIÓN

Actualmente la tecnología de la información juega un papel fundamental en la sociedad en la que vivimos, llegando incluso hasta el punto de la dependencia. Esto ha sido motivado por la era digital en la que esta sociedad se encuentra inmersa. La cantidad y diversidad de sistemas de información que se manejan diariamente es incontable: Librerías digitales, Sistemas de Información Geográfica (SIG)/Infraestructuras de Datos Espaciales (IDE), directorios y buscadores, etc. Todos ellos motivados por el deseo de hacer que la información esté disponible globalmente y pueda ser accesible para el mayor número de gente como sea posible en un entorno colaborativo. Por ello, resulta esencial organizar, publicitar y facilitar el acceso a dicha información. Y para que esto sea posible, es decir, para que un recurso sea encontrado como resultado de una búsqueda debemos ser capaces de describirlo según sus propiedades.

Es en este escenario heterogéneo dónde las descripciones de los recursos resultan críticas. Mediante los metadatos se pretende describir los recursos en base a sus propiedades, características y el contexto en el que el recurso toma sentido. Esto permite el descubrimiento, la indexación y la catalogación de los recursos de acuerdo a sus características (tipo de datos, contenido, origen, calidad, fecha de creación, etc.) y su contexto, para posteriormente poder ser encontrados [1]. Por lo tanto, los metadatos resultan ser la pieza clave de cualquier sistema de información [2].

La generación y/o creación de metadatos ha sido identificada como una tarea tediosa y poco gratificante, siendo necesario dedicar gran cantidad de tiempo y recursos, tanto económicos como humanos [3][4]. Además de resultar una tarea pesada, la compilación manual de metadatos, supone una fuente de errores por parte del creador [5]. En consecuencia, el objetivo de este trabajo es investigar nuevas técnicas y metodologías para generar la mayoría de las descripciones de metadatos de forma automática. De modo que se puedan describir de forma completa y veraz los recursos para posteriormente poder ser publicados y conseguir así facilitar a los usuarios el acceso a los mismos.

Cuando un usuario se encuentra frente a un recurso desconocido en términos de formato, lo primero que tiene que hacer es examinar y extraer la mayor cantidad de información posible del recurso en sí mismo, de su contexto y, obviamente, de su contenido. Pero esto no siempre es fácil, dado que la extracción automática de metadatos implica el conocimiento de las estructuras internas de los formatos de almacenamiento de datos utilizados por los recursos geográficos [4]. Este proceso normalmente lleva a cabo una correspondencia entre las características extraídas de cada formato y los distintos elementos de metadatos descritos por alguno o algunos de los estándares existentes (Dublin Core, ISO19115, etc.). Sin embargo, el gran número de formatos de datos existentes para los recursos geográficos hace muy difícil que una sola aplicación pueda manejar todos ellos. Un enfoque alternativo es el desarrollo de soluciones integradas y flexibles basadas en la reutilización de librerías, herramientas o componentes que son capaces de leer múltiples formatos para extraer la información de metadatos.

Si aparte de describir los recursos geográficos, se pretende generalizar el proceso de extracción de metadatos para cualquier tipo de recurso multimedia, entonces el problema se agrava, dado que el número de posibles formatos con los que se va a tener que tratar aumenta considerablemente. Por esa razón, se busca una plataforma que permita acceder a los recursos heterogéneos y obtener información de una manera homogénea.

DISCUSIÓN SOBRE HERRAMIENTAS Y PLATAFORMAS DE ACCESO A DATOS Y METADATOS

Como se ha mencionado en la introducción, la necesidad de acceder y de obtener información de tantos formatos como sea posible motivó un estudio que analiza y evalúa varias plataformas comunes que proporcionan acceso a información geográfica, así como varias soluciones de código abierto para la extracción de metadatos.

El objetivo es obtener descripciones de recursos basadas en la extracción de metadatos, por lo que las herramientas de extracción de metadatos deben ser consideradas en primer lugar. Analizando las capacidades de generación automática de metadatos de CatMDEdit [6], se puede ver que proporciona extracción de metadatos para varios de los formatos geográficos soportados [7]. Sin embargo, como se pretende generalizar el proceso de extracción de metadatos para cualquier tipo de recurso multimedia, se consideró la herramienta Apache Tika [8] como una solución que se adecua a los objetivos marcados. Actualmente, Apache Tika no incluye soporte para formatos geoespaciales, sin embargo tiene una arquitectura extensible que permite añadir nuevos formatos de datos. Como la extensibilidad es un requisito fundamental en este enfoque para que se pueda dar soporte a tantos tipos y formatos de recursos como sea posible, Apache Tika fue la herramienta de extracción de metadatos seleccionada.

Como paso previo a la extracción de los metadatos, la nueva solución tenía que ser capaz de acceder e interpretar los formatos de datos. A continuación, se discute sobre las plataformas de acceso a datos geográficos analizadas, que son, junto con Apache Tika, el otro componente de la solución integrada. La primera plataforma que se analizó fue GeoTools [9], las primeras pruebas de extracción de metadatos atrajeron inicialmente la atención como una buena solución, sin embargo, posteriormente resultó complicado ampliar la gama de formatos de recursos soportados como otras plataformas permiten, y que además ya dan soporte a más formatos. Por su parte, la capa de acceso a datos (DAL) de gvSIG [10] tiene como objetivo proporcionar a gvSIG una capa de abstracción que permite al núcleo de la aplicación operar de forma homogénea con diferentes fuentes de datos y formatos. Aunque DAL es conceptualmente compatible con la plataforma que se estaba buscando, todavía se encontraba en las primeras etapas de desarrollo. Las librerías GDAL/OGR [11] proporcionan acceso a una gran cantidad de formatos geográficos ráster y vectoriales, con un bajo nivel de abstracción. Por lo tanto, serían un buen punto de partida si se hubiese deseado empezar a desarrollar una plataforma común para acceder a datos geoespaciales. Pero mediante una solución con un nivel de abstracción mayor se puede facilitar el trabajo y ahorrar una gran cantidad de tiempo y esfuerzo. Por último, el proyecto OSGeo FDO [12] posibilita el acceso a diversas fuentes de datos geoespaciales a través de un mecanismo común. Soporta una gran variedad de fuentes de datos, incluyendo formatos de archivos, bases de datos y servicios geoespaciales. Se consideró que FDO puede ofrecer la funcionalidad deseada y es compatible con casi todos los formatos de información geográfica conocidos, por lo que fue la plataforma de acceso a datos seleccionada.

En resumen, el enfoque que se adoptó para acceder, interpretar y extraer los metadatos de los recursos geográficos y no geográficos combina los beneficios de dos herramientas independientes pero complementarias: Apache Tika y OSGeo FDO. A continuación se verá en mayor detalle cómo estas dos herramientas pueden trabajar juntas.

INTEGRACIÓN DE APACHE TIKA Y OSGEO FDO

Como se ha mencionado en la introducción, el objetivo de este trabajo es conseguir una herramienta que permita el acceso y la extracción de metadatos de una gran variedad de recursos. La aproximación que se considera aquí es la combinación de los proyectos Apache Tika y OSGeo FDO en una solución integrada que ofrece los beneficios de ambos proyectos. Mediante esta integración, se obtiene una poderosa herramienta de extracción de metadatos para gran variedad de tipos de recursos multimedia, con especial énfasis en recursos geoespaciales.

Apache Tika tiene una arquitectura extensible basada en el concepto de puntos de conexión o proveedores de datos, es lo que ellos llaman *parsers*. En resumen, un proveedor de datos maneja ciertos formatos de datos, lo que en términos de programación significa que un proveedor de datos es un *parser* dedicado a ciertos formatos. Esencialmente, Apache Tika se estructura en varios proveedores de datos para tener acceso a múltiples formatos de datos. Por lo tanto, Apache Tika se puede ampliar con nuevos proveedores de datos según sea necesario.

En particular, la solución planteada consistió en la adición de uno o más proveedores de datos basados en la API de FDO para que los múltiples formatos que soporta FDO pudieran ser soportados por la arquitectura de Apache Tika. Esto implica que todos los formatos de datos geográficos soportados por FDO fueron automáticamente integrados y accesibles a través de Apache Tika. Entonces, los clientes ganan acceso a una serie de nuevos tipos de recursos en cuanto los proveedores de datos nuevos se agregan.

Este tipo de arquitecturas cuyos proveedores de datos son independientes y extensibles presentan dos ventajas claras: reutilización y escalabilidad. El primero se refiere a la capacidad de reutilización de librerías actuales especializadas para construir los proveedores de datos. Aparte de algunas excepciones notables (por ejemplo GML), los formatos de datos por lo general sufren pocas y lentas modificaciones durante su ciclo de vida (por ejemplo XML, Atom o shapefile) dado que la mayoría de ellos están bajo el control de organismos de normalización que regulan la evolución y las modificaciones futuras de dichos estándares. Por lo tanto, envolver las librerías y herramientas estables existentes como proveedores de datos basados en Tika alivia notablemente el esfuerzo de crearlos y mantenerlos desde cero.

El segundo beneficio significa que el mecanismo extensible basado en la adición de los proveedores de datos permite escalar de forma adecuada la solución para múltiples escenarios y necesidades. Las relaciones de uno-a-muchos se pueden producir cuando un formato de datos cuenta con soporte de varios proveedores de datos. Sin embargo, en la práctica, finalmente sólo un proveedor de datos se encargará de un formato de datos establecido explícitamente en un archivo de configuración. Entonces, cada proveedor de datos es disjunto en términos de formatos de datos soportados, lo que supone que los proveedores de datos independientes puedan ser fácilmente agregados, eliminados o modificados, según sea necesario.

Como se ilustra en la Figura 1 los proveedores de datos, si es necesario, se basan en librerías Java que proporcionan la funcionalidad necesaria para procesar los formatos de datos específicos. Por ejemplo, el PDFParser procesa los recursos en formato PDF utilizando la librería Apache PDFBox[1]. O el HTMLParser que soporta virtualmente cualquier tipo de código HTML que se encuentra en la web utilizando la biblioteca TagSoup[2].

[1] http://pdfbox.apache.org
[2] http://home.ccil.org/~cowan/XML/tagsoup

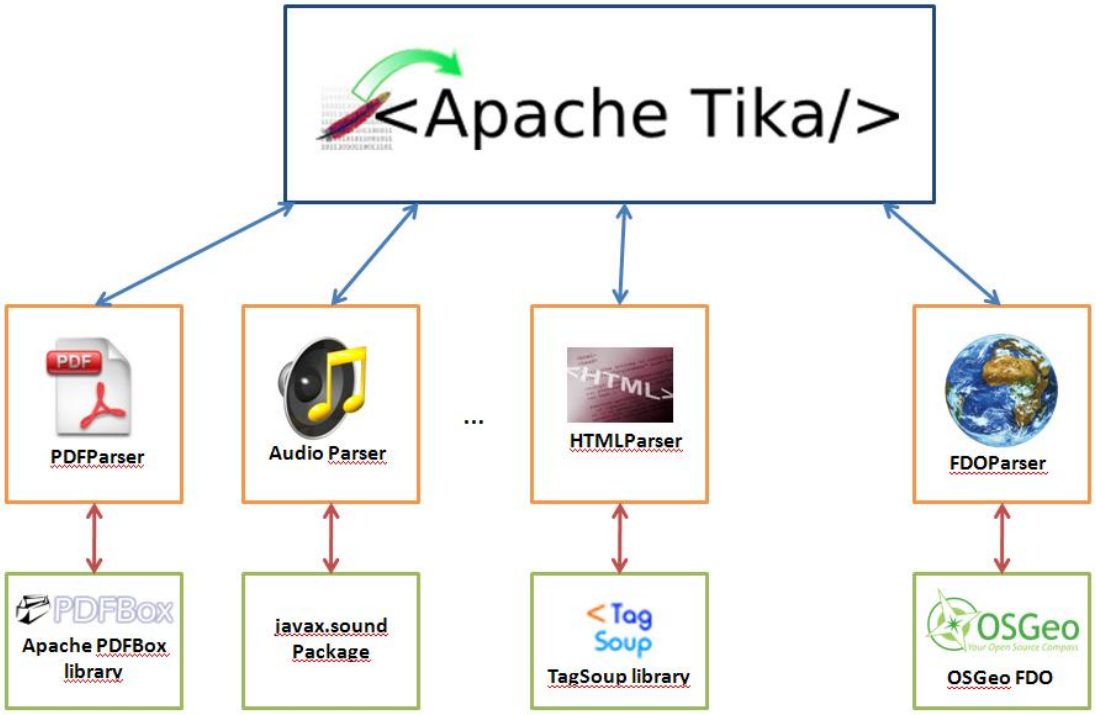

Figura 1: *Integración de Apache Tika y OSGeo FDO.*

Aunque conceptualmente la integración de Apache Tika y OSGeo FDO que se propuso puede parecer sencilla, a nivel técnico, sin embargo, el problema encontrado es que ambos proyectos utilizan diferentes lenguajes de programación, Java y C++/.Net respectivamente.

La solución que se suele elegir en este tipo de situaciones es el uso de Java Native Interface (JNI)[3]. JNI es un *framework* de programación que permite que un programa escrito en Java ejecutado en la máquina virtual java (JVM) pueda interactuar con programas escritos en otros lenguajes, como C o C++. Para facilitar la creación de código JNI, ya que puede ser una tarea difícil y tediosa, existen herramientas de desarrollo como el *Simplified Wrapper and Interface Generator* (SWIG)[4] que ayuda a los programadores a generar los enlaces entre proyectos Java y C++. SWIG es un compilador de interfaces que conecta programas escritos en C y C++ con varios lenguajes de programación de alto nivel, como Java. SWIG se usa normalmente para generar el código necesario para acceder a las interfaces C/C++ desde el lenguaje de programación deseado, de modo que podremos invocar la funcionalidad del código C/C++.

Tras mucho esfuerzo intentando generar los enlaces para la API de OSGeo FDO, el puente entre Java y C++ no se consiguió ni mediante JNI ni mediante SWIG. Quizá el problema más grande fuera la complejidad de las estructuras de datos necesarias para representar e intercambiar información geográfica. La solución adaptada pasó por el desarrollo de un servidor HTTP en C# mediante el cual se accede a la funcionalidad de OSGeo FDO a través de interfaces HTTP.

[3] http://java.sun.com/javase/6/docs/technotes/guides/jni/index.html
[4] http://www.swig.org

RESULTADOS

El prototipo funcional que integra Apache Tika y OSGeo FDO permite la extracción de metadatos de una amplia gama de formatos de recursos multimedia. En concreto, esta herramienta soporta los tipos de recursos inicialmente soportados por OSGeo FDO (más de 150 formatos de IG)[5] y los tipos de recursos que soporta Apache Tika (más de 50 formatos multimedia)[6]. Por lo que la solución integrada es compatible con más de 200 formatos de recursos multimedia en su conjunto.
El nivel de detalle de las descripciones que esta herramienta proporciona para cada tipo de recurso depende del *parser* usado para analizarlo y en última instancia del propio recurso. Por ejemplo, la cantidad de metadatos extraídos por los *parsers* ya incluidos en Apache Tika depende directamente del tipo de recurso analizado. Las figuras 2 y 3 muestran las descripciones de metadatos extraídas de un documento de texto en el formato de Microsoft Word[7] y de un archivo de audio en formato MP3[8] respectivamente.

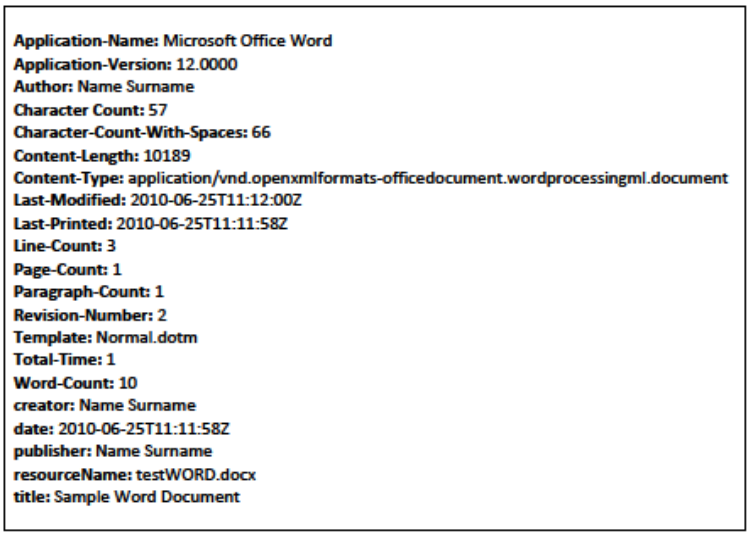

Figura 2: *Metadatos extraídos de un documento MSWord por Apache Tika.*

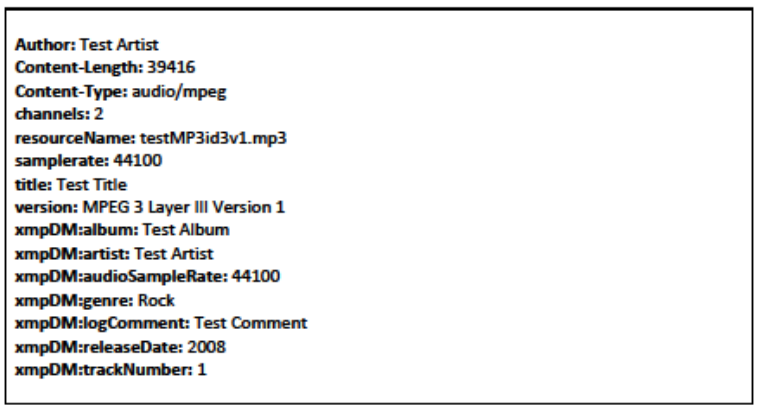

Figura 3: *Metadatos extraídos de un archivo de audio MP3 por Apache Tika.*

[5] http://fdo.osgeo.org/OSProviderOverviews.html
[6] http://tika.apache.org/0.8/formats.html
[7] http://office.microsoft.com/en-us/word
[8] http://es.wikipedia.org/wiki/MP3

El nuevo *parser* de OSGeo FDO permite analizar los recursos geoespaciales independientemente del lugar donde se almacenan y extraer descripciones uniformes de metadatos, independientemente del formato de datos del recurso. La Tabla 1 recoge un conjunto seleccionado de los descriptores de metadatos que se extraen a través del *parser* de OSGeo FDO. Estas etiquetas son descriptores comunes, independientemente del formato de los recursos. Sin embargo, las características específicas de cada formato pueden ser también capturadas.

Tabla 1: Resumen de los metadatos extraídos desde la API de FDO

Etiqueta	Significado
Resource	Inicia la descripción de un recurso
FormatName	Formato en el que se encuentra el recurso analizado
ResourceType	Tipo de recurso
Provider	Proveedor de FDO utilizado para analizar el recurso
Source	Origen de los datos
ResourceName	Nombre del recurso
ConnectionString	Cadena de conexión al recurso
dateStamp	Fecha de la creación de los metadatos
keywords	Palabras clave en la descripción del recurso
SpatialContexts	Descripción de los contextos espaciales del recurso. Incluyendo nombre, sistema de coordenadas, extensión…
Schemas	Descripción de los esquemas de datos del recurso. Incluyendo nombre, atributos, *features*…
SchemaAttributes	Descripción de los atributos del esquema. Incluyendo nombre y valor de los mismos.
FeatureClasses	Descripción de los *features* del recurso. Incluyendo nombre, restricciones, sus propiedades…
BaseIdentityProperties y Properties	Descripción de las propiedades de cada *feature*. Incluyendo nombre, tipo y diferente información dependiente del tipo, como tamaño, valor por defecto, precisión...

Además de estas etiquetas, el *parser* de OSGeo FDO también ha añadido a las descripciones de los servicios basados en estándares OGC[9] soportados toda la información procedente de un *GetCapabilities*. Cabe destacar que esta puede ser la principal fuente de información para este tipo de recursos, siempre y cuando sus responsables hayan dedicado cierto esfuerzo en completar sus metadatos.

En la solución también se ha incluido un módulo de configuración para especificar a priori algunas etiquetas de metadatos en un archivo XML que se incluirán automáticamente en todas las descripciones de metadatos. Por lo tanto, este módulo permite incluir información dependiente del contexto en cada descripción de metadatos de forma fácil y configurable. Los metadatos basados en el contexto pueden incluir información tan relevante como el autor del conjunto de datos o la empresa a la que pertenecen, entre otros.

[9] http://www.opengeospatial.org

CONCLUSIONES

Teniendo en mente el objetivo final de facilitar al usuario el acceso a los recursos, el primer paso para conseguirlo es la descripción de los recursos, en este caso mediante la generación automática de metadatos, para su posterior catalogación o indexación que permita organizar los recursos y finalmente publicarlos para poder ser encontrados.

Tras estudiar varias posibles soluciones, en este trabajo se han integrado los proyectos FDO y Apache Tika como capa de acceso a datos que posibilita la extracción de metadatos que permitan describir todo tipo de recursos multimedia. Observando los resultados conseguidos se puede concluir que la nueva plataforma da soporte para describir recursos multimedia de una gran variedad de formatos e incluso servicios, especialmente de formatos de IG. Por lo tanto, se considera que se ha cumplido el objetivo inicial de este trabajo, esto es lograr una plataforma que nos permita obtener información y acceder de forma lo más homogénea posible a recursos heterogéneos.

Actualmente, resulta muy complicado conseguir un sistema que permita la descripción de recursos totalmente autónomo, pues siempre será necesaria la participación del usuario para introducir o por lo menos validar los campos de metadatos menos intuitivos, hay que tener en cuenta que no todos los datos son fáciles de averiguar, por ejemplo el resumen o el título. Deberemos empezar por rellenar los campos básicos de descubrimiento de forma que se puedan ejecutar búsquedas mínimas con éxito, por ejemplo, en un catálogo. Más tarde podremos dedicar esfuerzos a completar rigurosamente el metadato. Es preferible tener todos los metadatos “a medias” que “atascarse” intentando rellenar exhaustivamente uno de ellos.

En este sentido, las descripciones conseguidas para los recursos de información geográfica en base a los metadatos extraídos de forma automática parecen ser bastante completas. Si a esto se le suma la participación del usuario a la hora de incluir más metadatos, ya sean como metadatos relativos al contexto preconfigurados o rellenando a mano los metadatos menos intuitivos, se puede conseguir un sistema que facilite en gran medida las rutinarias y poco motivadoras labores de los creadores de metadatos, reduciendo además los errores que se producen al escribir directamente los metadatos.

Finalmente, se considera esencial impulsar la investigación en todos los campos relacionados con la generación y gestión de metadatos dado el papel clave que estos juegan en cualquier sistema de información. Estos metadatos permiten indexar y catalogar los recursos de una forma más exacta en los sistemas de información y, en consecuencia, aumentan la capacidad de proporcionar resultados más relevantes y exactos a las búsquedas realizadas por los usuarios. Con todo esto, se consigue mejorar la accesibilidad a la información, el objetivo principal.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el proyecto “España Virtual” (ref. CENIT 2008-1030) a través del Instituto Geográfico Nacional (IGN) de España.

REFERENCIAS

[1] Smits PC and Friiss-Christensen A (2007) Resource Discovery in a European Spatial Data Infrastructure. IEEE Transactions on Knowledge and Data Engineering 19(1), pp. 85-95.

[2] Nogueras-Iso J, Zarazaga-Soria FJ, Muro-Medrano PR (2005) Geographic Information Metadata for Spatial Data Infrastructures: Resources, Interoperability and Information Retrieval. Springer, 264 p., ISBN: 978-3-540-24464-6.

[3] Tolosana-Calasanz R, Álvarez-Robles J, Lacasta J, Nogueras-Iso J, Muro-Medrano P and Zarazaga-Soria F (2006) On the Problem of Identifying the Quality of Geographic Metadata. Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science 4172, Springer. Berlin / Heidelberg, pp. 232-243.

[4] Manso MA, Nogueras-Iso J, Bernabé MA and Zarazaga-Soria FJ (2004) Automatic Metadata Extraction from Geographic Information, in: papers presented at the AGILE 2004 Conference, May 1st, 2004, Heraklion, Greece.

[5] Manso MA and Bernabé MA (2009) Metadatos implícitos de la información geográfica: caracterización del coste temporal y de los tipos y tasas de errores en la compilación manual. GeoFocus 9, pp. 317-336.

[6] IGN (2010) CatMDEdit OpenSource Project. Instituto Geográfico Nacional (IGN Spain) and University of Zaragoza. http://catmdedit.sourceforge.net, Último acceso 03.2011.

[7] Grupo de Sistemas de Información Avanzados (IAAA) de la Universidad de Zaragoza (2010) CatMDEdit User Manual v4.5. http://iaaa.cps.unizar.es/software/index.php/CatMDEdit_English_user_manual, Último acceso 03.2011.

[8] Apache Software Foundation (2010) Apache Tika: a content analysis toolkit. http://tika.apache.org, Último acceso 03.2011.

[9] Turton I (2008) GeoTools. En: G. B. Hall, M. G. Leahy (eds.), Open Source Approaches in Spatial Data Handling. Springer, pp. 153-169.

[10] Anguix A, Díaz L (2008) gvSIG: A GIS desktop solution for an open SDI. Journal of Geography and Regional Planning Vol. 1(3), May, 2008, pp. 041-048.

[11] Warmerdam F (2008) The Geospatial Data Abstraction Library. In: G. B. Hall, M. G. Leahy (eds.), Open Source Approaches in Spatial Data Handling. Springer, pp. 87-104.

[12] OSGeo FDO (2010) Feature Data Objects (FDO) Data Access Technology. Open Source Geospatial Foundation. http://fdo.osgeo.org, Último acceso 03.2011.